Multiple Linear Regression - Interactions

Ryan Miller

Introduction

- One of the beauties of multiple regression is its capacity to isolate the distinct effects for two (or more) predictors of a single outcome
  - However, sometimes predictors do not make independent contributions towards the outcome, and instead work *synergistically* to produce an outcome

Categorical-Quantitative Interactions

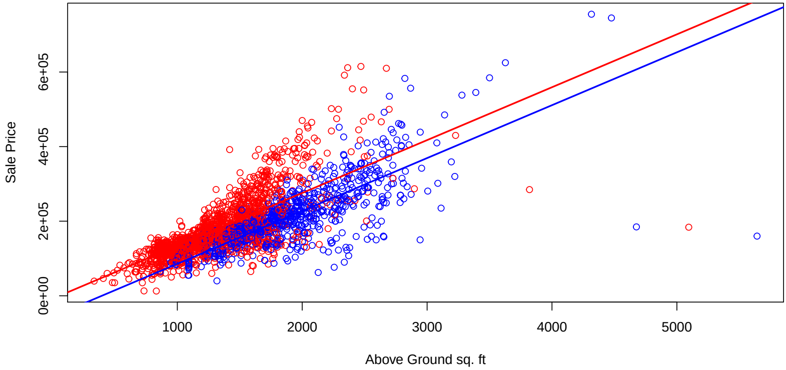
- Let's look at the relationship between above ground living area and sale price for 1Story and 2Story homes in the Ames Housing dataset
  - Recall the coefficient for the dummy variable "House.Style2Story" in is negative, how did we interpret this?

```
m1 <- lm(SalePrice ~ House.Style + Gr.Liv.Area, data = ah)
m1$coefficients
```

```
##       (Intercept) House.Style2Story       Gr.Liv.Area
##        -7931.5874       -48161.2973          141.7925
```

Categorical-Quantitative Interactions

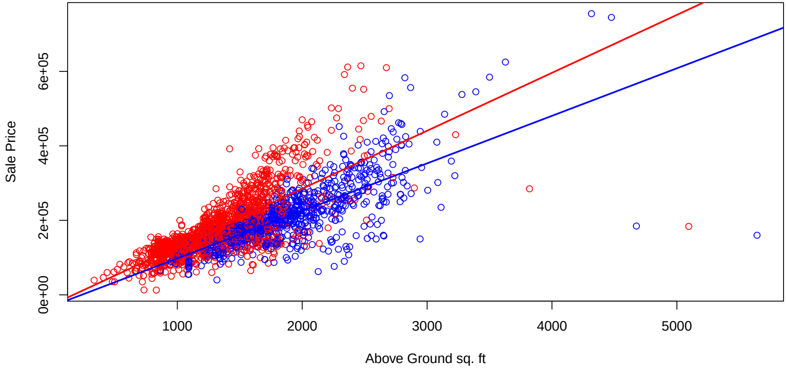
- If two homes are the same size, our model predicts the "2Story" home will be cheaper
  - Further, this model estimates a single, adjusted slope for "Gr.Liv.Area" (regardless of whether a home is "1Story" or "2Story")



Categorical-Quantitative Interactions

- Now let's consider a model with an **interaction** between "House.Style" and "Gr.Liv.Area"

```
m2 <- lm(SalePrice ~ House.Style + Gr.Liv.Area + House.Style*Gr.Liv.Area, data = ah)
```



Categorical-Quantitative Interactions

- The interaction term allows for different slopes depending upon the value of the "House.Style" dummy variable
  - When the dummy variable takes on a value of 0, 155.6 is the slope (in the "Gr.Liv.Area" dimension)
  - When the dummy variable takes on a value of 1, 155.6 - 27.9 = 127.7 is the slope

```
m2$coefficients
```

```
##                 (Intercept)            House.Style2Story
##                 -26041.68634                 -3758.60579
##                  Gr.Liv.Area House.Style2Story:Gr.Liv.Area
##                    155.55156                   -27.92985
```

Quantitative-Quantitative Interactions

- The same general concepts apply to interactions between two quantitative variables, though interpretation can be more difficult
  - The coefficients of the model `SalePrice ~ Year.Built + Gr.Liv.Area + Year.Built*Gr.Liv.Area` are shown below
  - Why is the coefficient of `Gr.Liv.Area` negative in this model?

```
m4 <- lm(SalePrice ~ Year.Built + Gr.Liv.Area + Year.Built*Gr.Liv.Area, data = ah)
m4$coefficients
```

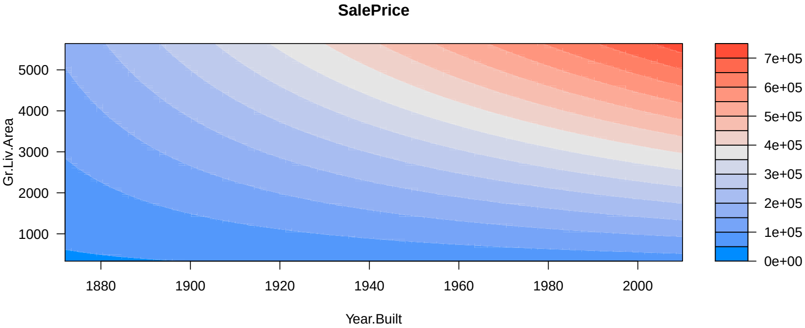
```
##            (Intercept)             Year.Built            Gr.Liv.Area
##          29162.4122638              3.7255782          -1334.7456056
## Year.Built:Gr.Liv.Area
##              0.7250126
```

Quantitative-Quantitative Interactions

- The estimated slope in the "Gr.Liv.Area" dimension will be different for each value of "Year.Built"
  - For a home built in the year 0 (nonsensical), the effect of "Gr.Liv.Area" is -1334
  - For a home built in 1900, the effect of "Gr.Liv.Area" is -1334 + 0.725*1900 = 43.5
  - For a home built in 2010, the effect is -1334 + 0.725*2010 = 123.3

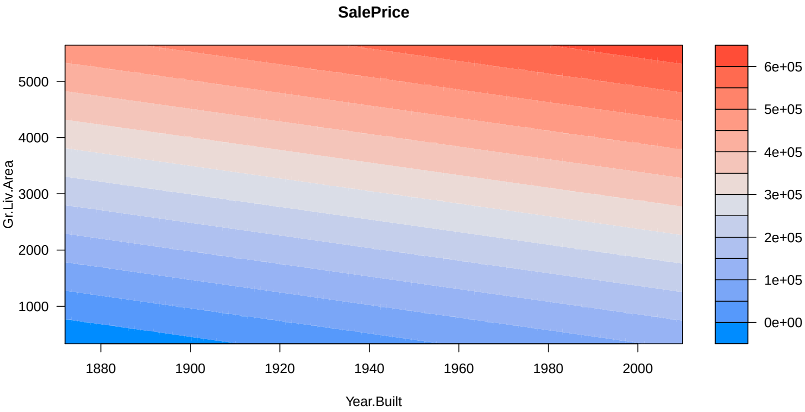
Quantitative-Quantitative Interactions

- Since there are now infinitely many slopes to consider, visualizing the model's predictions is arguably a more useful approach
  - This plot emphasizes that high values in both "Year.Built" and "Gr.Liv.Area" work in tandem to produce a high sale price



Visualizing the Model w/o an Intercation

- When there's no interaction, we can see the slope is constant in each dimension



Categorical-Categorical Interactions

- A final scenario to consider is an interaction between two categorical predictors
  - This is equivalent to giving each cell in the two-way table it's own effect

```
mc <- lm(SalePrice ~ House.Style + Foundation + House.Style*Foundation, data = ah)
mc$coefficients
```

```
##                         (Intercept)                  House.Style2Story
##                           99151.573                          40621.374
##                    FoundationCBlock                    FoundationPConc
##                           47752.687                         133014.245
##                      FoundationSlab                    FoundationStone
##                            4778.733                          16848.427
##                      FoundationWood House.Style2Story:FoundationCBlock
##                          102848.427                         -30670.578
##   House.Style2Story:FoundationPConc   House.Style2Story:FoundationSlab
##                          -36807.409                          -5132.251
##   House.Style2Story:FoundationStone   House.Style2Story:FoundationWood
##                            7238.501                           7378.626
```

Categorical-Categorical Interactions

- For example, in our data the 2Story PConc homes have a mean sale price of $235,980
  - This is expressed by our model as: 99152 (intercept) + 40621 (main effect of 2Story) + 133014 (main effect of PConc) - 36807 (interaction of 2Story and PConc)
- How could you use the model to find the mean sale price of 1Story Slab homes?

House.Style	Foundation	mean
1Story	BrkTil	99151.57
1Story	CBlock	146904.26
1Story	PConc	232165.82
1Story	Slab	103930.31
1Story	Stone	116000.00
1Story	Wood	202000.00
2Story	BrkTil	139772.95
2Story	CBlock	156855.06
2Story	PConc	235979.78
2Story	Slab	139419.43
2Story	Stone	163859.88
2Story	Wood	250000.00

Closing Remarks

- Interactions are one way of making linear regression models more flexible, but in doing so they can sometimes open up a can of worms
  - Even a relatively tame modeling application involving only 10 predictors results in $\binom{10}{2} = 45$ possible interactions to consider
- In most applications, statisticians will only consider interactions if there is sufficient rationale for doing so
  - This is usually based upon the scientific context of the modeling application and the current knowledge in that field